NLP methods for crypto price prediction

A. Kipriyanov G. Kuzmin

ICEF

December 18, 2022

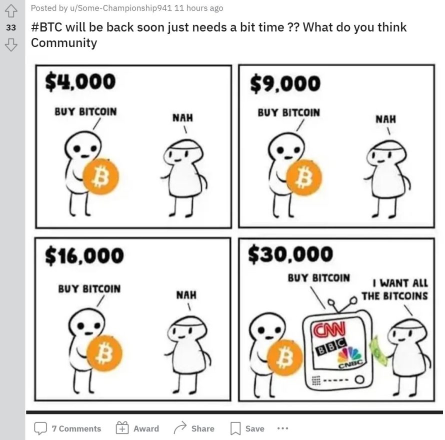
Idea of the research

- Cryptocurrency – people's money
- Easy access by millions of people
- Social networks (Facebook[1], Twitter, Reddit, etc)
- Pure speculative value creation, no underlying asset
- Retrieve sentiment ⇒ forecast the direction of the return
- Compare with more traditional econometric and ML models

[1] parent company 'Meta' is recognized as an extremist organization and banned on the territory of the Russian Federation

Data (1/2)

Typical Reddit posts:

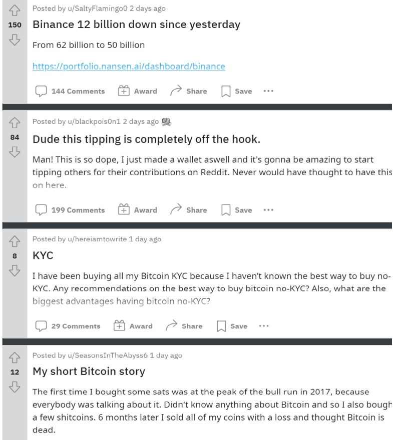


33

Posted by u/Some-Championship941 11 hours ago

#BTC will be back soon just needs a bit time ?? What do you think Community

7 Comments Award Share Save ···



150

Posted by u/SaltyFlamingo0 2 days ago

Binance 12 billion down since yesterday

From 62 billion to 50 billion

https://portfolio.nansen.ai/dashboard/binance

144 Comments Award Share Save ···

84

Posted by u/blackpois0n1 2 days ago

Dude this tipping is completely off the hook.

Man! This is so dope, I just made a wallet aswell and it's gonna be amazing to start tipping others for their contributions on Reddit. Never would have thought to have this on here.

199 Comments Award Share Save ···

8

Posted by u/hereiamtowrite 1 day ago

KYC

I have been buying all my Bitcoin KYC because I haven't known the best way to buy no-KYC. Any recommendations on the best way to buy bitcoin no-KYC? Also, what are the biggest advantages having bitcoin no-KYC?

29 Comments Award Share Save ···

12

Posted by u/SeasonsInTheAbyss6 1 day ago

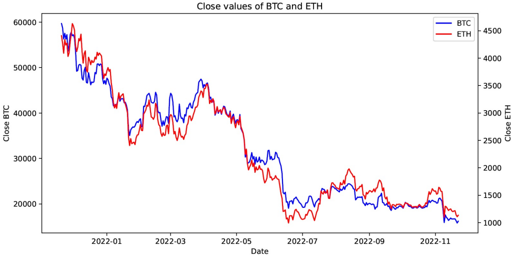
My short Bitcoin story

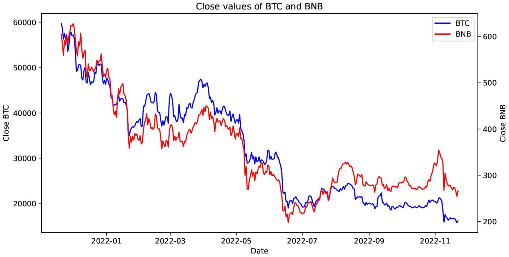
The first time I bought some sats was at the peak of the bull run in 2017, because everybody was talking about it. Didn't know anything about Bitcoin and so I also bough a few shitcoins. 6 months later I sold all of my coins with a loss and thought Bitcoin is dead.

We collect Score, Subreddit, Title, Post Text, ID, Total Comments, Date

Data (2/2)

Yahoo!Finance: Bitcoin, Ethereum, USDT, USDC, Binance Coin and Dogecoin





NLP (1/3)

Marking Data

negative (-1)	neutral (0)	positive (1)
return $<$ -2%	return $\in$ [-2%; +2%]	return $>$ +2%

TF-IDF

$$tf = \frac{\text{\# of the word in the text file}}{\text{total \# of all words in the text file}}$$

$$idf = log(\frac{\text{\# of other text files}}{\text{\# of text files with this word}})$$

$$tf - idf = tf \times idf$$

- Support Vector Machine (SVM) to perform classification
- Does not count semantic similarity between words

NLP (2/3)

Word2Vec

- Attaching semantic numeric vector to each word and obtaining aggregated one for the post
- Finding similar semantic vectors via cosine similarity

KNN prediction algorithm

For post $i = 1$ to $M + V$ (M - train sample, V - test sample):

Step 1: for $m = 1$ to M: if $i \notin M$, estimate their cosine similarity (d_{im})

Step 2: Select KNN(highest d) and take initial sentiments (logreturn)

Step 3:

$$Sent_t = \frac{\sum\limits_{\{m \in KNN\}} d_{im}^3 \times Initial\ Sent}{\sum\limits_{\{m \in KNN\}} d_{im}^3}$$

Step 4: $s_t = \dfrac{\sum\limits_{\{m \in KNN\}} d_{im}^3}{K}$

Step 5:

$$s_{pred} = \frac{\sum\limits_{t=1}^{N} s_t}{N}; \quad Sent_{pred} = \frac{\sum\limits_{t=1}^{N} Sent_t \times s_t}{\sum\limits_{t=1}^{N} s_t}$$

NLP (3/3)

TF-IDF + SVM	Word2Vec + KNN
Train 80%, Test 20%	2 days to predict next day
Expanding window	Rolling window

F-beta=0.7027
MSE=1.3067
	F-beta=0.6071
	MSE=1.1429

Limitations and Improvements

- **TF-IDF**:
  - Add additional factors, such as posts' length, binary for emoji-es and presence of video materials
  - Other alternative classification techniques could be used such as logit, Random Forests
- **Word2Vec**
  - Extremely computationally extensive, that is why, we should check the performance with the training data larger than posts within two days

Potential applications:

Constructing trading algorithm, based on predicted marking: (*positive* (1) → *Buy*, *neutral* (0) → *Hold*, *negative* (-1) → *Sell*)

Competing methodologies (1/2)

- Polynomial regression
  - Lags of BTC, ETH, BNB and DOGE Returns, Lags of BTC Volume and Range, Score and Total Comments
  - Train-validation-test split: 6-2-2
  - CV $\Rightarrow$ power $= 1$
  - MSE $= 0.9185$
- Random Forest
  - Lags of BTC, ETH, BNB and DOGE Returns, Lags of BTC Volume and Range, Score and Total Comments
  - Train-test split: 8-2
  - 10-fold CV $\Rightarrow$ # trees $= 520$
  - **MSE** $= 0.0337$
- GARCH (1,1)
  - Return of BTC
  - Train-test split: 8-2
  - MSE $= 0.0464$

Competing methodologies (2/2)

- ANN
  - forecasts of Polynomial regression, Random Forest and GARCH (1,1)
  - 2 hidden layers, 4096 neurons each plus a concatenation layer
  - MSE $= 0.03507$
- ANN+
  - forecasts of all models plus lags of data
  - 2 hidden layers, 4096 neurons each plus a concatenation layer
  - MSE $= 1.3806$
- RNN
  - Return of BTC
  - Simple RNN layer
  - **MSE** $= 0.0248$

Improvements and Further Research

- Increase the data set
- Try more sophisticated ANN (eg. LSTM)
- Parce pictures and determine the sentiment using the body of the post

Improvements and Further Research

- Increase the data set
- Try more sophisticated ANN (eg. LSTM)
- Parce pictures and determine the sentiment using the body of the post